How To Talk to Robots: Evidence from User Studies on Human-Robot Communication

Petra Gieselmann
Interactive Systems Lab
University of Karlsruhe
Germany
petra@ira.uka.de

Prisca Stenneken
Physiological and Clinical Psycholgoy
Cath. University Eichstätt-Ingolstadt
Germany
stenneken.ku-eichstaett@gmx.de

Abstract

Talking to robots is an upcoming research field where one of the biggest challenges are misunderstandings and problematic situations: Dialogues are error-prone and errors and misunderstandings often result in error spirals from which the user can hardly escape. Therefore, mechanisms for error avoidance and error recovery are essential. By means of a data-driven analysis, we evaluated the reasons for errors within different testing conditions in human-robot communication and classified all the errors according to their reasons. For the main types of errors, we implemented mechanisms to avoid them. In addition, we developed an error correction detection module which helps the user to correct problems. Therefore, we are developing a new generation strategy which includes detecting problematic situations, helping the user and avoiding giving the same information to the user several times. Furthermore, we evaluate the influence of the user strategy on the communicative success and on the occurrence of errors within human-robot communication. In this way, we can increase the user satisfaction and have more successful dialogues within human-robot communication.

1 Introduction

We developed a household robot which helps users in the kitchen (Gieselmann et al., 2003). It can get something from somewhere, set the table, switch on or off lamps or air conditioners, put something somewhere, tell the user what is in the fridge, tell some recipes, etc. The user can interact with the robot in natural language and tell it what to do. A first semantico-syntactic grammar has been developed and we now enhance this dialogue grammar by means of user tests and data collections.

Since the real robot consists of many different components, such as the speech recognizer, the gesture recognizer, the dialogue manager, the motion component, etc., we decided to restrict the user tests for the beginning to the dialogue management component. This means that we do not use a real robot to accomplish the tasks, but only a text-based interface where the dialogue manager informs the user what the robot is doing. In this way, we can skip problems resulting from other components and can focus on understanding and dialogue problems. We are aware of the fact that the findings cannot be directly applied to spoken communication with the real robot. However, this text-based paradigm was used for a first systematic investigation and is transferred to spoken robot communication in future studies.

In this paper, we discuss two methods how to improve human-robot communication: By analysing human-robot dialogues and avoiding the most important problems and on the other hand by changing the communicative strategy of the user. The second section deals with related work. Section three explains our household robot, the dialogue system and its particular characteristics. The fourth section is about user tests within different testing conditions which results in an error classification. Section five addresses the question whether communicative strategies affect the human-robot communication both in the subjective evaluation by the users and in the objectively measurable task success. Section six gives a

conclusion and an outlook on future work.

2 Related Work

2.1 Errors in Man-Machine Dialogues

Most of the research about errors within man-machine dialogues deal with speech recognition errors: Some researchers evaluate methods for dialogue state adaptation to the language model to improve speech recognition (Xu and Rudnicky, 2000; Gorrell, 2003). Work on hyperarticulation concludes that speakers change the way they are speaking when facing errors in principle so that the language model has to be adapted (Stifelman, 1993; Hirschberg et al., 2004). Also Choularton et al. and also Stifelman are looking for general strategies on error recognition and repair to prepare the speech recognizer for the special needs of error communication (Choularton and Dale, 2004; Stifelman, 1993).

Furthermore, Schegloff et al. came up with a model which describes the mechanisms the dialogue partners use to handle errors in human-human dialogue (Schegloff et al., 1977). Also, within conversation analysis dialogues are evaluated concerning the rules and procedures how an interaction takes place (Sacks et al., 1974). These insights from human-human communication are essential for a natural human-robot communication.

However, the present study concentrates on semantic errors and classify them according to their reasons. For every error class, we develop methods to avoid it. Furthermore, we examine repair dialogues and their similarity to human-human repair dialogues in order to be able to perform efficient error handling strategies so that it will be easier for the user to correct errors which could not be avoided.

2.2 Effects of the User Strategy on Dialogue Success

In the field of humanoid robots and human-robot interaction the researchers concentrate on questions such as how to design the robot as similar as possible to a human regarding its outer appearance as well as its communicative behaviour (Breazeal, 1999; Billard and Mataric, 2000; Dautenhahn and Billard, 1999). In contrast, the present study concentrates on the human user and his communication strategies. This in turn would shape the expectations on how the dialogue should work and how errors could be avoided by another user strategy.

Figure 1: Our Household Robot

Furthermore, different evaluation methodologies of dialogue systems exist, starting from methodologies using the notion of a reference answer (Hirschmann et al., 1990) to the most prominent approach for dialogue system evaluation which is Paradise (Walker et al., 1997) which uses a general performance function covering different measures such as user performance, number of turns, task success, repair ratio, etc. In the present study, objective measures were calculated from the participants' responses and success measures were assessed after each block in form of a questionnaire in order to get a deeper insight in the relationship between subjective and objective measures of success.

3 Our Household Robot

3.1 The Dialogue Manager

For dialogue management we use the TAPAS dialogue tools collection (Holzapfel, 2005) which is based on the approaches of the language and domain independent dialogue manager ARIADNE (Denecke, 2002). This dialogue manager is specifically tailored for rapid prototyping. Possibilities to evaluate the dialogue state and general input and output mechanisms are already implemented which are applied in our application. We developed the domain and language dependent components, such as an ontology, a specification of the dialogue goals, a data base, a context-free

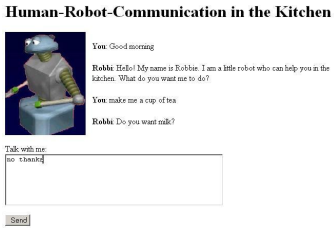

Figure 2: The web-based Interface of our Humanoid Robot

grammar and generation templates.

The dialogue manager uses typed feature structures (Carpenter, 1992) to represent semantic input and discourse information. At first, the user utterance is parsed by means of a context-free grammar which is enhanced by information from the ontology defining all the objects, tasks and properties about which the user can talk. In our scenario, this ontology consists of all the objects available in the kitchen and their properties and all the actions the robot can do. The parse tree is then converted into a semantic representation and added to the current discourse. If all the necessary information to accomplish a goal is available in discourse, the dialogue system calls the corresponding service. But if some information is still missing, the dialogue manager generates clarification questions to the user. This is realized by means of generation templates which are responsible for generating spoken output.

3.2 Rapid prototype

We developed a rapid prototype system. This system includes about 32 tasks the robot can accomplish and more than 100 ontology concepts. Ontology concepts can be objects, actions or properties of these objects or actions. By means of this prototype we started user tests and continue to develop new versions of the grammar and domain model. The rapid prototype of our dialogue component is integrated in the robot (cf. figure 1) and also accessible via the internet for the web-based tests (cf. figure 2).

4 Analysis of Human-Robot Dialogues

4.1 Different Testing Conditions

As mentioned by Dybkjaer and Bernsen (Dybkjr and Bernsen, 2000), predefined tasks covered in a user test will not necessarily be representative of the tasks real users would expect a system to cover. In addition, scenarios in user tests should not prime users on how to interact with the system which can only be avoided in a user test without predefined tasks or in a general user questionnaire. On the other hand, such a free exploration is much more complicated for the user and can be very frustrating, if the system does not understand the user intention. Therefore, we rely on two different testing conditions:

User tests with predefined tasks: Every user got five predefined tasks to accomplish by means of the robot. Since the tasks are given, it is easier for the user, but we do not get any information on the tasks a user really needs a robot for.

User tests without predefined tasks: The users were just told that they bought a new household robot which can support them in the household. They can freely explore and interact with the robot. This situation is much more realistic, but at the some time much harder for the user because he does not know what the robot can do in detail.

In addition, we had two different testing conditions: Web-based user tests (see Figure 2) which have the advantage that lots of users all over the world can participate whenever they like to (Schmidt, 1997; Reips, 2002) and also multimodal user tests with the robot (see Figure 1) to see how the user can get along with the real robot. The tests with the web-interface are of course different from the ones with the real robot, but within the web tests we can also use more dialogue capabilities concerning tasks the robot cannot accomplish until now.

4.2 Experimental Details and Results

We defined all the user turns which could not be transformed to the correct semantics by the

	Robot	Web-based
With Tasks	22.57%	49.94%
Without Tasks	57.03%	50.93%

Table 1: Turn Error Rates Within Different Testing Conditions.

dialogue system as *errors* so that the turn error rate gives the rate of error turns on the whole number of user turns. As expected, the turn error rate for tests with tasks is lower than without tasks (cf. Table 1) given the fact that the user has less clues what to say. Especially the tests with predefined tasks with the robot results in much less errors which might be due to the fact that these tasks were easier than in the web-based test and that the users could watch the robot interacting.

Nevertheless, within all the testing conditions, we can find the same error classes according to the following reasons for failure:

- **New Syntactic and Semantic Concepts**: New Formulations, New Objects, New Goals, Metacommunication
- **Ellipsis & Anaphora**: Elliptical Utterances, Anaphora, Missing Context
- **Concatenated Utterances**
- **Input Problems**: Punctuation & Digits, Background Noise, Grammatically Wrong Utterances

In addition, the rates for the error classes are very similar so that most of the errors can be found in the area of new syntactic and semantic concepts, secondmost errors are input errors, thirdmost ellipsis and the fewest errors belong to the class of concatenated utterances.

Since the manual integration of new concepts is very time and cost-intensive, we developed a mechanism for dynamic vocabulary extension with data from the internet (Gieselmann and Waibel, 2006). In addition, we implemented mechanisms to deal with ellipsis and anaphora (Gieselmann, 2005) and handle complex user utterances. To resolve metacommunication, we grouped all the user utterances dealing with metacommunication according to the user intention:

- **Clarification Questions** from the user: The user wants to know, whether the robot understood him, what the robot is doing, etc.
- **Repair** of a user utterance: The user corrects the preceding utterance of the robot explicitly or implicitly.
- **Test** of the Robot: The user tests the abilities of the robot by giving instructions for tasks the robot can probably not accomplish; also insults are in this category.

Clarification questions from the user and tests of the robot indicate that the user does not know what the robot can do, has no idea on how to go on and what to say. Therefore, we implemented communication strategies so that the robot explains its capabilities to the users and help them in the case of problems. Different factors can indicate communication problems, such as that the user utterance is inconsistent with the current discourse, it cannot be completely parsed, it does not meet the system expectations, the user says the same utterance several times. These factors leads to an increase in error correction necessity and let the robot finally initiate a clarification dialog to help the user.

5 Influence of the User Strategy on the Communicative Success

5.1 Experimental Details

To evaluate the influence of the user strategy on the communicative success and the occurrence of errors, we conducted a web-based experiment with two different instructions for each participant:

- "Child instruction": The users were asked to talk to the robot in the same way as they would do to a little child.
- "Non-child instruction": The users got no detailed instruction on how to talk to the robot.

Each participant got predefined tasks. During the user interaction with the system, we measured the objective success per user by means of the turn error rate, the number of successfully accomplished tasks and the number of

user turns necessary to accomplish resp. abort a task. After the participants had finished the task set under each instruction, they filled in a short user questionnaire about their general impression of the system and their experience during the experiment.

5.2 Results and Discussion

The effects of the instruction child vs. non-child are reflected in both qualitative and quantitative measures. Within quantitative measures, the instruction affected above all the mean utterance length, ie. number of words per user utterance. Participants had a numerically lower mean utterance length with instruction child (mean = 5.02) as compared to the non-child instruction (mean = 5.64). Interestingly, the effect of smaller mean utterance lengths in the child instruction occurs predominantly when the child instruction is given in the second block (the modulatory effect of the order of the instruction was marginally significant, $p = .053$). This might be due to the fact that participants who got the child instruction in the first block continued with this strategy also in the second block, irrespective of the instruction. This fact is also reported by some participants in the post-test questionnaires. Also within qualitative measures, about half of the participants reported to use short, simple sentences within the child instruction.

Pairwise comparisons were performed for possible effects of the instruction on subjective or objective measures of communicative success. For all variables, the effects of the instruction were non-significant, although we found a tendency towards more user satisfaction in the child instruction. This might be due to the fact that the present instructions were given rather implicitly and left some space for individual interpretations.

As expected, when comparing subjective and objective measures, a significant correlation was observed for the subjective measure "willingness to use the system again" and the objective measure "overall number of accomplished tasks" (p-value smaller than .05). Even though all other correlations did not reach significance, the numerical tendencies imply that the more tasks are accomplished, the higher the ratings are for subjective variables.

Findings from analyses of the user answers in free text also suggest that we have a rather strong influence of the participants' general attitude towards robots which has a more dominant effect on the task success than the instruction. Since the conversation style of the user seems to be affected to a larger extent by the general attitude, future studies might address the question, how a dialogue system has to be designed to find out different user attitudes, support them and their different characteristics to improve the communication and avoid errors.

6 Conclusion and Outlook

We used a date-driven method to evaluate the reasons for errors in human-robot communication and implemented the following strategies to avoid resp. deal with them:

- dynamic extension of linguistic resources
- anaphora resolution
- handling complex as well as elliptical utterances
- meta communication

We evaluated the influence of the user strategy on the communicative success and found out that even though the user strategy had qualitative and quantitative effects on the communicative behavior, it was not systematically related to the communicative success in objective and subjective measures. However, the general attitude of the user towards robots has a more dominant effect on the task success than the instructed user strategy.

Future studies could further address the question, whether these findings are also true for extended grammars and tests with the real robot instead of the web interface.

References

A. Billard and M. J. Mataric. 2000. A biologically inspired robotic model for learning by imitation. *Proceedings of the 4th conference on Autonomous Agents.*

C. Breazeal. 1999. Robot in society: Friend or appliance? *Proceedings of the Agents99 workshop on emotion-based agent architectures.*

B. Carpenter. 1992. *The Logic of Typed Feature Structures*. Cambridge University Press.

S. Choularton and R. Dale. 2004. User responses to speech recognition errors: Consistency of behaviour across domains. *Proceedings of the Tenth Australian International Conference on Speech Science and Technology*.

K. Dautenhahn and A. Billard. 1999. Bringing up robots or the psychology of socially intelligent robots: from theory to implementation. *Proceedings of the 3rd conference on Autonomous Agents*.

M. Denecke. 2002. Rapid prototyping for spoken dialogue systems. *Proceedings of the 19th International Conference on Computational Linguistics*.

L. Dybkjr and N.O. Bernsen. 2000. Usability issues in spoken language dialogue systems. *Kuppevelt, J. v., Heid, U. and Kamp, H. (Eds.): Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, Natural Language Engineering*, 6:243–272.

P. Gieselmann and A. Waibel. 2006. Dynamic extension of a grammar-based dialogue system: Constructing an all-recipes knowing robot. *To Appear in: Proceedings of the International Conference on Spoken Language Processing (ICSLP 06)*.

P. Gieselmann, C. Fügen, H. Holzapfel, T. Schaaf, and A. Waibel. 2003. Towards multimodal communication with a household robot. *Proceedings of the Third IEEE International Conference on Humanoid Robots (Humanoids)*.

P. Gieselmann. 2005. Reference resolution mechanisms in dialogue management. *Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue (CATALOG)*.

G. Gorrell. 2003. Recognition error handling in spoken dialogue systems. *Proceedings of the 2nd International Conference on Mobile and Ubiquitous Multimedia*.

J. Hirschberg, D. Litman, and M. Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43.

L. Hirschmann, D. A.Dahl, D. P. McKay, L. M. Norton, and M. C. Linebarger. 1990. Beyond class a: A proposal for automatic evaluation of discourse. *Proceedings of the Speech and Natural Language Workshop*, pages 109–113.

H. Holzapfel. 2005. Towards development of multilingual spoken dialogue systems. *Proceedings of the 2nd Language and Technology Conference*.

U.-D. Reips. 2002. Standards for internet-based experimenting. *Experimental Psychology*, 49(4).

H. Sacks, E. Schegloff, and G. Jefferson. 1974. A simple system for the organization of turntaking im conversation. *Language*, 50(4):696–735.

E. Schegloff, G. Jefferson, and H. Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language 53*.

W. C. Schmidt. 1997. World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments & Computers*, 29(2).

L. J. Stifelman. 1993. User repairs of speech recognition errors: An intonational analysis. *Technical Report, Speech Research Group, MIT Media Lab*.

M. A. Walker, D. Litman, C. A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280.

W. Xu and A. Rudnicky. 2000. Language modeling for dialog system. *Proceedings of the International Conference of Speech and Signal Processing (ICSLP'00)*.